U.S. Department
of Transportation
**Bureau of
Transportation Statistics**

BTS Guide to Good Statistical Practice

September 2002

1. Introduction

Quality of data has many faces. Primarily, it has to be *relevant* to its users. Relevance is an outcome that is achieved through a series of steps starting with a planning process that link user needs to data requirements. It continues through acquisition of data that is *accurate* in measuring what it was designed to measure and produced in a *timely* manner. Finally, the data must be made *accessible* and *easy to interpret* for the users. In a more global sense. data systems also need to be *complete* and *comparable*. The creation of data that addresses all of the facets of quality is a unified effort of all of the development phases from the initial data system objectives, through system design, collection, processing, and dissemination to the users. These sequential phases are like links in a chain. The sufficiency of each phase must be maintained to achieve relevance. This document is intended to help management, and data system sponsors achieve relevance through that sequential process.

1.1 Legislative Background

The 1991 Intermodal Surface Transportation Efficiency Act (ISTEA) created the Bureau of Transportation Statistics (BTS) within the Department of Transportation (DOT). Among other things, it made BTS responsible for: "issuing guidelines for the collection of information by the Department of Transportation required for statistics … in order to ensure that such information is accurate, reliable, relevant, and in a form that permits systematic analysis." (49 U.S.C. 111 (c)(3))

A parallel requirement for developing guidelines emerged in the Paperwork Reduction Act of 1995. It tasked the Office of Management and Budget (OMB) to "develop and oversee the implementation of Government wide policy, principles, and guidelines concerning statistical collection procedures and methods; statistical data classification; statistical information presentation and dissemination; timely release of statistical data; and such statistical data sources as may be required for the administration of federal programs." (44 U.S.C. 3504 (e)(3))

Lastly, the Consolidated Appropriations Act of 2001, section 515, elaborated on the Paperwork Reduction Act, requiring OMB to issue guidelines ensuring the quality of disseminated information by 9/30/2001 and each federal agency to issue guidelines by 9/30/2002.

1.2 OMB Guidelines for Ensuring Information Quality

On 28 September 2001, OMB published a notice in the Federal Register (finalized as 67 FR 8452, February 22, 2002) that required agencies to issue guidelines for

ensuring and maximizing the quality of information disseminated by federal agencies.

As defined in the OMB guidance, quality consists of:

- Utility, i.e., the usefulness of information to intended users,
- Objectivity in presentation and in substance, and
- Integrity, i.e., the protection of information from unauthorized access or revision.

Agencies were required to develop guidelines, covering all information disseminated on or after October 1, 2002, regardless of format. Agencies were also required to develop a process for pre-dissemination review of information, an administrative mechanism allowing the public to request correction of information not complying with the guidelines, and an annual report to OMB indicating how the public requests were handled by the mechanism.

These guidelines incorporate the statistical aspects of the OMB guidelines as a baseline and elaborate on its recommendations to produce statistical guidelines adapted for the Department of Transportation.

1.3 Applicability

These guidelines apply to all statistical information that is disseminated on or after 1 October 2002 by agencies of the Department of Transportation (DOT) to the public using the "dissemination" definition in the OMB guidelines. That definition exempts a number of classes of information from these guidelines. Major types of exempted information are listed below. A more detailed list is provided in section IV of the DOT Information Dissemination Quality Guidelines, of which this document is a subsection.

- Information disseminated to a limited group of people and not to the public in general.
- Archival records that are inherently not "active."
- Materials that are part of an ad judicatory process.
- Hyperlinked information.
- Opinion offered by DOT staff in professional journals.

DOT disseminated data contain a lot of information provided by "third party sources" like the states, industry organizations, and other federal agencies. These guidelines apply to that disseminated data unless exempted for other reasons discussed above. However, DOT guidelines indicating design, collection, and processing methods do not apply to data acquisition steps performed by non-federal sources. Steps performed by federal sources outside DOT before proving the data to DOT will be governed by the agency's own guidelines in accordance with this legislation. For data provided to DOT by third party sources, these guidelines primarily emphasize disseminating information about data quality, the DOT processing methods, and analysis of the data provided to the users.

1.4 Types of DOT Statistical Data Collected

The recommendations within these guidelines apply to a wide range of data collection types. They include reporting systems, surveys, and special studies.

Reporting systems are set up to be automatic delivery of data into the data system. They collect incident information from government (federal, state, or local) and industry sources and periodic information on transportation flow and volume from government and industry. Incident data tend to cover all incidents (e.g., fatal accidents), though some data may be sampled due to its shear volume (e.g., highway injuries). Flow and volume is a mixture of 100% collection and sampled data. Surveys and special studies are more of an outreach form of data collection. Surveys and studies are usually conducted using some form of sampling.

Samples taken for any data collection may be selections of people or organizations from lists, samples of geographic areas or sections of highway, or samples of time segments.

1.5 Overview of the Statistical Guidelines

The quality guidelines for statistical information are based on structured planning (section 2), sound statistical methods (sections 3 and 4) and the principle of openness (sections 5 and 6). Structured planning maintains the link between user needs and data system design. Sound statistical methods produce information (data and analysis results) that conforms to that design. Openness ensures that users of statistical information can easily access and interpret the information.

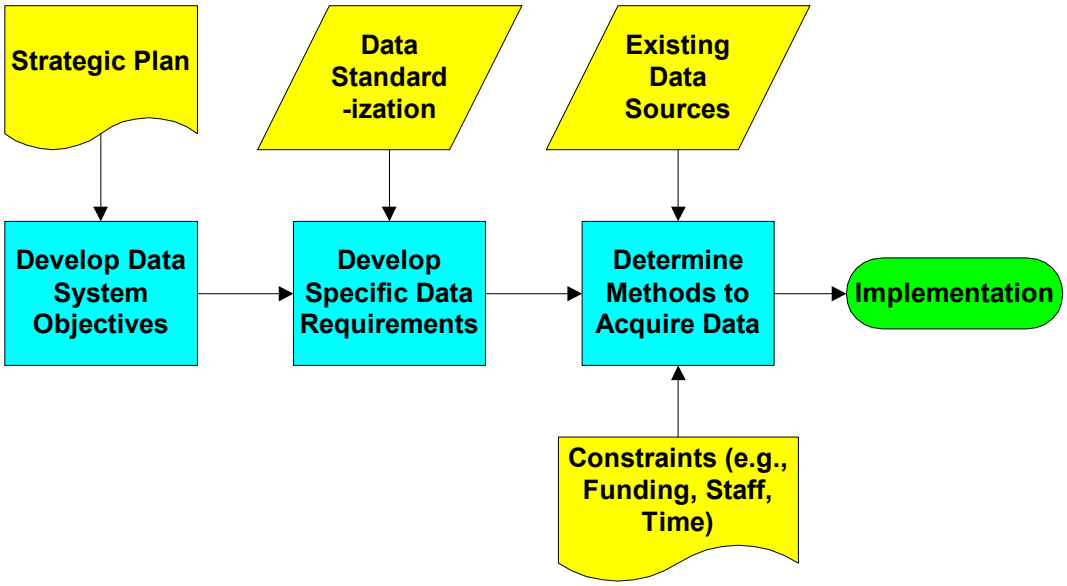
Each section begins with a statement of *principles*, which contain definitions, assumptions, and rules or concepts governing action. The *principles* are followed by *guidelines*, which are specific recommended actions with examples. Finally, each section concludes with *references*.

1.6 Statistical Guidelines Relationship to DOT's Information Dissemination Quality Guidelines

These statistical guidelines are a subset of the DOT Information Dissemination Quality Guidelines. Chapters 2 through 6 discussed above form section VI, paragraphs a – e in that document.

2. Planning Data Systems

A data system produced within a DOT agency is linked to that organization's strategic planning. The data are compiled to measure success toward a goal, satisfy an external user need (which should also be a goal), or used as a tool necessary to perform work toward a goal. Data system planning consists of three stages: development of objectives for the system, translation of those objectives into data requirements, and planning of the top-level methods that will be used to acquire the data.



2.1. Data System Objectives

Principles

- The "system sponsor" and "sponsoring organization" as used in these guidelines is the organizational entity whose strategic plan and budget will guide the creation of the data system. It is usually at the agency level.

- These guidelines assume that the sponsoring organization's strategic plan is current and contains all of its goals and objectives, including those relative to the creation of the data system.

- "Objectives" of the data system describe what federal programs and external users will accomplish with the information. They should be traceable to the strategic plan goals.

- System objectives in clear, specific terms, identifying data users and key questions to be answered by the data system, will help guide the system development to produce the results required.

- Just as strategic plans change over time, the objectives of the data system will need to change over time to meet new requirements.

- Users will benefit from knowing the objectives that guided the system design.

Guidelines

- Every data system objective should be traceable to the goals and objectives in the sponsoring organization's strategic plan.

> For example, NHTSA's primary goal is to improve traffic safety, so one initial objective for the Fatality Analysis Reporting System (FARS) could be to provide an overall measure of highway safety as an objective basis to evaluate the effectiveness of highway safety improvement efforts.

- The system sponsor should develop and update the data system objectives in partnership with critical users and stakeholders. The sponsor should have a system to regularly update the system as user needs change.

- The objectives should indicate each major need that will be fulfilled by the system and the data users associated with that need, and the key questions that will be answered by the data.

> For example, for the Highway Performance Monitoring System (HPMS), one of the objectives may be: to provide state and national level measures of the overall condition of the nation's public road systems as investment information for Congress, condition and performance information for the traveling public, and information necessary to make equitable apportionments of highway funds to the states.

- Objectives should include timeliness of the data.

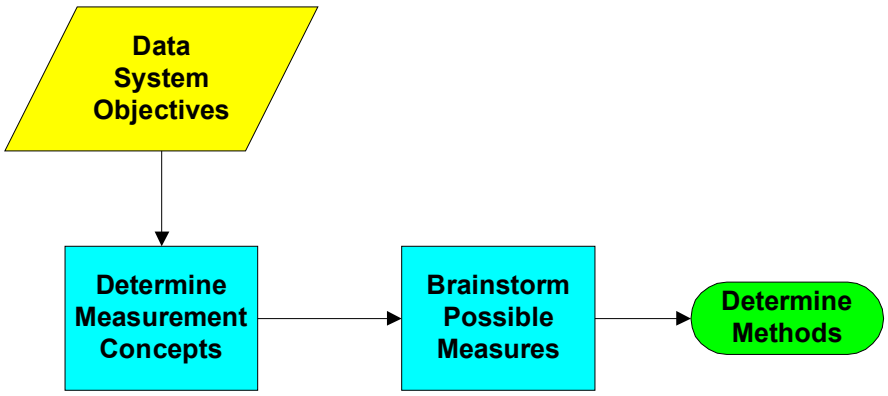
- The current data system objectives should be documented and clearly posted with the data, or with the disseminated output from the data.

- The updating system should be documented and include how user information is collected.

References

- Huang, K., Y.W. Lee, and R.Y. Wang. 1999. Quality Information and Knowledge. Saddle River, NJ: Prentice Hall.

2.2. Data Requirements



Principles

- A "measurement concept" is a characteristic of people, businesses, objects, or events (e.g., people or businesses in a city or state, cars or trains in the United States, actions at airports, incidents on highways).

> Examples: The level of success stopping illicit drug smuggling into the U.S. over maritime routes. The level of use of public transit in a metropolitan area.

- Before deciding on what data should be in a data system or how to acquire them, the data system objectives need to be linked to more specific "measurement concepts," from which data requirements will be derived.

> Example: For FARS, the objective "To provide an overall measure of highway safety" leads to the measurement concept of "The safety of people and pedestrians on the highways of the U.S."

- Measurement concepts related to objectives can be outcomes that change as objectives are achieved, outputs from agency accomplishments related to an objective, efficiency concepts, inputs, and quality of work.

- From the measurement concepts, data requirements are created for possible measurement of each measurement concept.

- Maintaining the link from data system objectives to measurement concepts to data requirements will help to ensure "relevance" of the data to users.

- In the data requirements, the use of standard names, variables, numerical units, and codes allow data comparisons across databases.

- Besides data that are directly related to strategic plans, additional data may be required for possible cause and effect analysis.

> For example, data collected for traffic accidents may include weather data for causal analysis.

Guidelines

- Each data system objective should have one or more "concepts" that need to be measured. Characteristics or attributes of the target group that are the focus of the objective should be covered by one or more measurement concepts.

> For HPMS, the objective "to provide a measure of highway road use" can lead to the measurement concept of "the annual volume of vehicles on state and interstate roads."

- The measurement concepts should be those characteristics which, when changing in a favorable way, indicate progress toward achievement of an objective.

> Note: Exceptions to this description are measures of magnitude, such as a total population or total vehicle miles traveled. These are "denominator measures" used to allow comparisons over time.

- Once the measurement concepts are chosen, develop data requirements needed to quantify them.

> Example: For HPMS, the measurement concept, "the annual volume of vehicles on state and interstate roads" can lead to a data requirement for state-level measures of annual vehicle-miles traveled accurate to within 10 percent at 80 percent confidence.

- There is usually more than one way to quantify a measurement concept. All reasonable measures should be considered without regard to source or availability of data. The final data choices will be made in the "methods" phase based on ease of acquisition, constraining factors (e.g., cost, time, legal factors), and accuracy of available data.

> Example: A concept of commercial airline travel "delay" can be measured as a percent of flights on-time in accordance with schedule, or a measure of average time a passenger must be in the airport including check in, security, and flight delay (feasibility of measure is not considered at this stage).

- The data requirements for each type of data should include required accuracy, timeliness, and completeness. The accuracy should be based on how the measure will be used for decision-making.

> Example: For FARS, the concept, "The safety of people and pedestrians on the highways of the U.S." can lead to data requirements for counts of fatalities, injuries, and motor vehicle crashes on U.S. highways and streets. The fatalities for a fiscal year should be as accurate as possible (100% data collection), available within three months after the end of the fiscal year, and as complete as possible. The injury counts in traffic accidents for the fiscal year totals should have a standard error of no more than 6 percent, be available within three months after the end of the fiscal year, and have an accident coverage rate of at least 90 percent.

- When selecting possible data, consider standardization with other databases. First, consider measures used for similar concepts in other DOT databases. Second, consider measures for similar concepts in databases outside DOT (e.g., The Census). Coding standards should be used where coding is used and made part of the data requirements. Such standardization leads to "coherence" across datasets.

> Examples: the North American Industry Classification System (NAICS) codes, the Federal Information Processing Standards (FIPS) for geographic codes (country, state, county, etc.), the Standard Occupation Codes (SOC), International Organization for Standardization (ISO) codes (money, countries, containers)

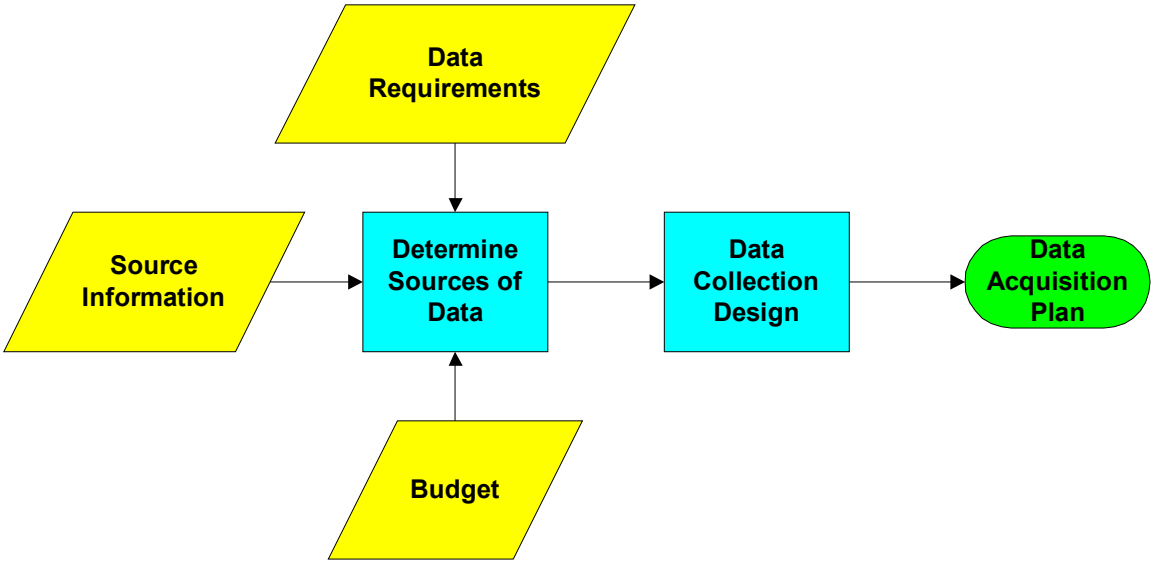
- The current data system measurement concepts and data requirements should be documented and clearly posted with the data.

References

- The Federal Information Processing Standards (FIPS) home page, http://www.itl.nist.gov/fipspubs/
- The North American Industry Classification System, http://www.census.gov/epcd/www/naics.html
- OMB Primer on Performance Management dated 2/28/1995
- American Association for Public Opinion Research. 1998. "Standard Definitions – Final Dispositions of Case Codes and Outcome Codes for RDD Telephone Surveys and In-Person Household Surveys." http://www.aapor.org/ethics/stddef.html.

2.3. Methods to Acquire Data

Given data requirements for a wide range of possible measures, the next phase is to consider the realities associated with gathering the data to construct estimates and perform analysis. After looking at the ease of data acquisition, complexity of possible acquisition approaches, budget restrictions, and time considerations, the list of possible measures is likely to be reduced to a more reasonable level. First, consider possible sources of data and then the process of acquiring it.



The more critical data needs invariably require greater accuracy. This in turn usually leads to a more complex data collection process. As the process gets more complex, there is no substitute for expertise. If the expertise for a complex design is not available in-house, consider acquiring the expertise by either contacting an agency that specializing in statistical data collection like the Bureau of Transportation Statistics or by getting contractual support.

2.4. Sources of Data

Principles

- A common arrangement in transportation is a reporting system in which the target group automatically sends data. Most of these are dictated by law or regulation. That limits the collection planning to working out the physical details.

> For example: 46 USC Chapter 61 specifies a marine casualty reporting system, while 46 CFR 4.05 specifies details.

- Use of existing data is by far the most efficient (i.e., cheapest) approach to data acquisition. Sources of existing data can be current data systems or administrative records.

- "Administrative records" are data that are created by government agencies to perform facilitative functions, but do not directly document the performance of mission functions (National Archives definition). In addition to providing a source for the data itself, administrative records may also provide information helpful in the design of the data collection system (e.g., sampling lists, stratification information).

> For example, state driver's license records, social security records, IRS records, boat registration records, mariner license records.

- Another method, less costly than developing a new data collection system, is to use existing data collections tailored to your needs. The sponsor of such a system may be willing to add additional data collection or otherwise alter the collection process to gather data for your needs.

> For example, the Bureau of Transportation Omnibus survey is a monthly transportation survey that will add questions related to transportation for special collections of data from several thousand households. This method could be used if this process is accurate enough for the data system needs.

- The "target group" is the group of all people, businesses, objects, or events about which information is required.

> For example, the following could be target groups: all active natural gas pipelines in the U.S. on a specific day, traffic accidents in FY2000 involving large trucks, empty seat-miles on the MARTA rail system in Atlanta on a given day, hazardous material incidents involving radioactive material in FY2001, mariners in distress on a given day, and all U.S. automobile drivers.

- One possible approach is to go directly to the "target group," either all of them (100%) or a sample of them. This would work with people or businesses.

- Another method frequently necessary with transportation data is the use of third party sources. Third party sources are people, businesses, or even government entities that have knowledge about the target group or collect information for other purposes, such as investigators, observers, or service providers (e.g., doctors).

> Examples: traffic observers, police observers, investigators, bus drivers counting passengers, state data collectors.

Guidelines

- Research whether government and private data gathering systems already have data that meet the data requirements. Consider surveys, reporting systems, and administrative records.

- If existing data meet some but not all of the data requirements, determine whether the existing data collection system can be altered to meet the data needs.

> For example, another agency may be willing to add to or alter their process in exchange for financial support.

- A primary consideration in whether to gather data from the target group or an indirect source is the access to those sources; all of those sources. A 100% data gathering would obviously need access to the entire target group. A sample approach will not include the entire target group, but all members should have a non-zero probability (and known) of selection, or the sampling will not necessarily be representative of the target group.

- Consider getting information directly from the target group (if they are people or businesses), having the target group observed (events as they occur), or getting information about the target group from another source (third party source discussed above).

- In some situations, the information desired is not directly available. In this case, consider collecting related information that can be used to derive or estimate the information required.

> For example: Collecting the number of people on and off a bus at each stop combined with a separate estimate of trip length between stops to estimate passenger miles.

- The choices made for sources and their connection to the data requirements should be documented and clearly posted with the data, or with disseminated output from the data.

References

- Electronic Records Work Group Report to the National Archives and Records Administration dated September 14, 1998.

2.5. Data Collection Design

Principles

- The design of data collection is one of the most critical phases in developing a data system. The accuracy of the data and of estimates derived from the data are heavily dependent upon the design of data collection.

> For example, the accuracy is dependent upon proper sample design, making use of sampling complexity to minimize variance. The data collection process itself will also determine the accuracy and completeness of the raw data.

- For large target groups, data collection from 100% of the target group is usually the most accurate approach, but is not always feasible due to cost, time, and other resource restrictions. It also is often far more accurate than the data requirements demand and can be a waste of resources.

- A "probability sample" is an efficient way to automatically select a data source representative of the target group with the accuracy determined by the size of the sample.

- When sampling people, businesses, and/or things, sampling lists (also known as frames) of the target group are required to select the sample. Availability of such lists is often a restriction to the method used in data collection.

- For most statistical situations, it is usually important to be able to estimate the variance along with estimating the mean or total.

- Sample designs should be based on established sampling theory, making use of multi-staging, stratification, and clustering to enhance efficiency and accuracy.

- Sample sizes should be determined based on the data requirements for key data, taking into account the sample design and missing data.

Guidelines

- If the target group is large, the data collection designer should use a probability sample, unless a 100% collection is required by law, necessitated by accuracy requirements, or turns out to be inexpensive (e.g., data readily available).

> For example, a system that collects data to estimate the total vehicle miles traveled (VMT) for a state of the U.S. cannot possibly collect 100 percent of all trips on every road, so a sampling approach is necessary. However, when it comes to collecting passenger miles for a large transit system, it may be possible with fare cards and computer systems to collect 100% of passenger miles.

- The sample design should give all members of the target group a non-zero (and known) probability of being represented in the sample.

> DANGER => Samples of convenience, such as collecting transportation counts at an opportune location, will produce data, but it will almost always be so biased as to be useless. Whereas, selecting locations using all possible locations in a sampling system will be statistically sound (with allowances due to correlations between locations).

- The design of any samples should be based on established sampling theory.

- Determine sample size using appropriate formulas to ensure data requirements for accuracy are met with adjustments for sample design and missing data. Use an appropriate random method to select sample according to the design.

- If some form of sampling is used, design the data collection to collect sufficient information to estimate the variance of each estimate to be produced.

- The collection design and its connection to the data requirements should be documented and clearly posted with the data, or with disseminated output from the data. The documentation should include references for the sampling theory used.

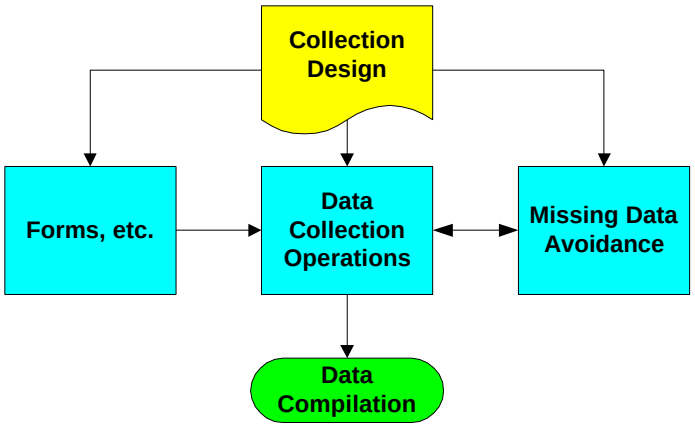
- If the data collection process uses sampling, a statistician or other sampling expert should develop or review the design.

References

- Cochran, William G., *Sampling Techniques* (3rd Ed.), New York: Wiley, 1977.

3. Collection of Data

Given the collection design, the next phase in the data acquisition process is the collection process itself. This collection process can be a one-time execution of a survey, a monthly (or other periodic) data collection, a continuous reporting of incident data, or a compilation of data already collected by one or more third parties. The physical details of carrying out the collection are critical to making the collection design a reality.



3.1 Data Collection Operations

Principles

- Forms, questionnaires, automated collection screens, and file layouts are the medium through which data are collected. They consist of sets of questions or annotated blanks on paper or computer that request information from data suppliers. They need to be designed to maximize communication to the data supplier.

- Data collection includes all the processes involved in carrying out the data collection design to acquire data. Data collection operations can have a high impact on the ultimate data quality, especially when they deviate from the design.

- The data collection method should be appropriate to the data complexity, collection size, data requirements, and amount of time available.

> For example, a reporting system will often primarily rely on the required reporting mechanism, with follow-up for missing data. Similarly, a large survey requiring a high response rate will often start off with a mail out, followed by telephone contact, and finally by a personal visit.

- Specific data collection environmental choices can significantly affect error introduced at the collection stage.

> For example, if the data collector is collecting as a collateral duty or is working in a uncomfortable environment, it may adversely affect the quality of the data collected. Also, if the data are particularly difficult to collect, it will affect the data quality.

- Conversion of data on paper to electronic form (e.g., key entry, scanning) introduces a certain amount of error which must be controlled.

- Third party sources of data introduce error in their collection processes.

- Computer-assisted information collection can result in more timely and accurate information. Initial development costs will be higher, and much more lead time will be required to develop, program, and test the data collection system. However, the data can be checked and corrected when originally entered, key-entry error is eliminated, and the lag between data collection and data availability is reduced.

- The use of sensors for data can significantly reduce error.

Guidelines

- Forms, screens, or file layouts used for data collection are clearly defined for data suppliers, with entries in a logical sequence, reasonable visual cues, and limited skip patterns. Instructions should help minimize missing data and response error.

- Computer assisted collection should be considered when the collection is repetitive over a long period of time making the gains in quality and data processing time worth the expense. Use of sensors (e.g., GPS, counters) should be considered to reduce error.

> Examples: Central telephone interviewing with computer screens and data entry by the interviewer. Handheld devices for entering train inspection data on-scene. Traffic counters with automatic upload to a central location.

- A status tracking system should be used to ensure that data are not lost in mailings, file transfers, or collection handling.

- Data entry of paper forms should have a verification system based on data accuracy requirements.

> For example, the verification samples of key entry forms can be based on an average outgoing quality limit for batches of forms. A somewhat more expensive approach would be 100 percent verification.

- Make the data collection as easy as possible for the collector.
- If interviewers or observers are used, a formal training process should be established to ensure proper procedures are followed.
- Data calculations and conversions at the collection level should be minimized.

> For example, if a bus driver is counting passengers, they should not be doing calculations such as summations. The driver should record the raw counts and calculations should be performed where they are less likely to result in mistakes.

- The collection operation procedures should be documented and clearly posted with the data, or with disseminated output from the data.

References

- Federal Committee on Statistical Methodology. 1983. *Approaches to Developing Questionnaires.* Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 10).
- Groves, R. 1989. *Survey Errors and Survey Costs*. New York, NY: Wiley, Chs. 10 & 11.

3.2 Missing Data Avoidance

Principles

- Some missing data occur in almost any data collection effort. Unit-level missing data occur when a report that should have been received is completely missing or is received and cannot be used (e.g., garbled data, missing key variables). Item-level missing data occur when data are missing for one or more items in an otherwise complete report.

> For example, for an incident report for a hazardous material spill, unit-level missing data occur if the report was never sent in. It would also occur if it was sent in, but all entries were obliterated. Item-level missing data would occur if the report was complete, except it did not indicate the quantity spilled.

- The extent of unit-level missing data can sometimes be difficult to determine. If a report should be sent in whenever a certain kind of incident occurs, then non-reporters can only be identified if crosschecked with other data sources. On the other hand, if companies are required to send in periodic reports, the previous period may provide a list of the expected reporters for the current periods.

> Both can also be true for item-level missing data. For example, in a travel survey asking for trips made, forgotten trips would not necessarily be known.

- Some form of missing data follow-up will dramatically reduce the incident of both unit-level and item-level missing data.

> For example, a system to recontact the data source can be used, especially when critical data are left out. A series of recontacts may be used for unit nonresponse. Incident reporting systems can use some form of cross-check with other data sources to detect when incidents occur, but are not reported.

- When data are supplied by a third-party data collector, some initial data check and follow-up for missing data will dramatically reduce the incident of missing data.

Guidelines

- Data collection programs should be conducted in a manner that is likely to produce high rates of response.

- All data collection programs require some follow-up of missing reports and data items, even if the data are provided by third-party sources.

> For example, for surveys and periodic reports, it is easy to tell what is missing at any stage and institute some form of contact (e.g., mail out, telephone contact, or personal visit) to fill in the missing data. For incident reports, it is a little more difficult, as a missing report may not be obvious.

- For incident reporting systems where missing reports may not be easily tracked, some form of checking system should exist to reduce missing reports.

- When collecting data from units of varying sizes (e.g., companies), the follow-up scheme should be prioritized, re-contacting larger reporters first, possibly at the risk of missing smaller reporters.

- For missing data items the data collection sponsor should distinguish between: critical items like items legally required or otherwise important items (e.g., items used to measure DOT or agency performance).

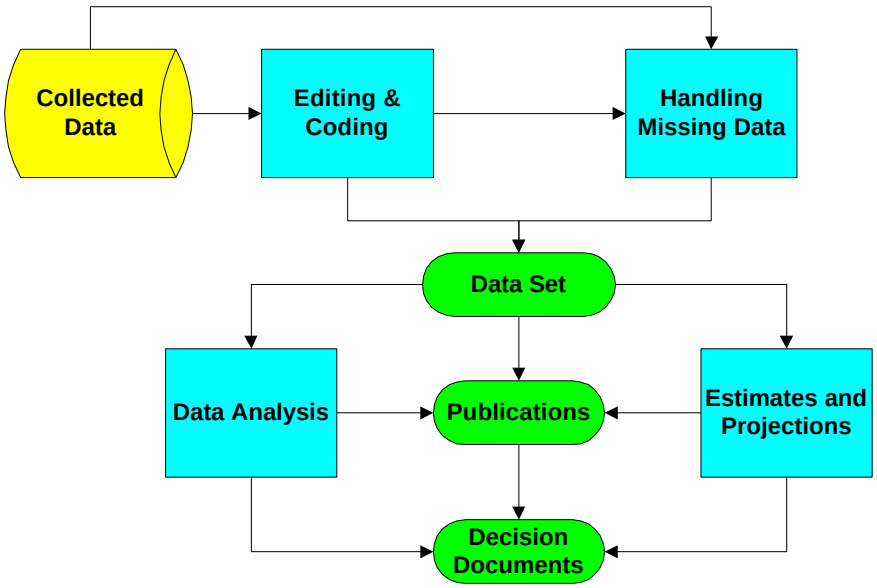
- The missing data avoidance procedures should be documented and clearly posted with the data, or with disseminated output from the data.

References

- Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York, NY: Wiley.

4. Processing Data

Given the collected data converted to electronic form, some "processing" is necessary to mitigate the obvious errors, and some analysis is usually necessary to convert data into useful information for decision documents, publications, and postings for the Internet.



4.1 Data Editing and Coding

Principles

- Data editing is the application of checks that identify missing, invalid, duplicate, inconsistent entries, or otherwise point to data records that are potentially in error.
- Typical data editing includes range checks, validity checks, consistency checks (comparing answers to related questions), and checks for duplicate records.
- For numerical data, "outliers" are not necessarily bad data. They should be examined for possible correction, rather than systematically deleted.

> **Note**: By "examine" we mean you can check the original forms, compare data items with each other for consistency, and/or follow-up with the original source, all to see if the data are accurate or error has been introduced.

- Editing is a final inspection-correction method. It is almost always necessary, but data quality is better achieved much earlier in the process through clarity of definitions, forms design, data collection procedures, etc.

- "Statistical edits" are methods for examining statistical properties of the data to detect more subtle errors.

> For example, examining distributions of variables for outliers, distribution anomalies, scatter plots of two related variables, and examining ratios of related variables or one variable over time.

- Coding is the process of adding codes to the data set as additional information or converting existing information into a more useful form. Some codes indicate information about the collection. Other codes are conversions of data, such as text data, into a form more useful for data analysis.

> For example, a code is usually added to indicate the "outcome" of each case. If there were multiple follow-up phases, the code may indicate in which phase the result was collected. Codes will also be added to indicate editing and missing data actions taken. Text entries are often coded to facilitate analysis. So, a text entry asking for a free form entry of a person's occupation may be coded with a standard code to facilitate analysis.

- Many coding schemes have been standardized.

> Examples: the North American Industry Classification System (NAICS) codes, the Federal Information Processing Standards (FIPS) for geographic codes (country, state, county, etc.), the Standard Occupation Codes (SOC).

Guidelines

- An editing system should be applied to every data collection and to third-party data to reduce obvious error in the data. A minimum editing process should include range checks, validity checks, checks for duplicate entries, and consistency checks. Consider statistical edits (see definition in principles) to detect more subtle errors.

> Examples of edits: If a data element has five categories numbered from 1 to 5, an answer of 8 should be edited to delete the 8 and flag it as a missing data value. Range checks should be applied to numerical values (e.g., income should not be negative). Rules should be created to deal with inconsistency (e.g., if dates are given for a train accident and the accident date is before the

> departure date, the rule would say how to deal with it). Data records should be examined for obvious duplicates.

- A recommended approach to editing is to make as many editing decisions as possible in advance and automate it. Reliance on manual intervention in editing should be minimized, since it may introduce human error.

- Avoid the overuse of outlier edits. Outliers can be very informative for analysis. Over-editing can lead to severe biases resulting from fitting data to implicit models imposed by the edits.

> Rapid industry changes could be missed if an agency follows an overly restrictive editing regimen that rejects large changes.

- Some method should be used to allow after-the-fact identification of edits. One method is to add a separate field containing an edit code (i.e., a "flag"). Another is to keep "version" files, though this provides less information to the users.

- To avoid quality problems from analyst coding and spelling problems, text information to be used for data analysis should be coded using a standard coding scheme (e.g., NAICS, SOC, and FIPS discussed above). Retain the text information for troubleshooting.

- The editing and coding process should clearly identify missing values on the data file. The method of identifying missing values should be clearly described in the file documentation. Special consideration should be given to files that will be directly manipulated by analysts or users. Blanks or zeros used to indicate missing data have historically caused confusion. Also, using a coding to identify the reason for the missing data will facilitate missing data analysis.

- The editing and coding process and editing statistics should be documented and clearly posted with the data, or with disseminated output from the data.

References

- Little, R. and P. Smith, "Editing and Imputation for Qualitative Survey Data," *Journal if the American Statistical Association*, Vol. 82, No. 397, pp. 58-68.

4.2 Handling Missing Data

Principles

- Untreated, missing data can introduce serious error into estimates. Frequently, there is a correlation between the characteristics of those missing and variables to be estimated, resulting in biased estimates. For this reason, it is best to employ adjustments and imputation to mitigate this damage.

- Without weight adjustments or imputation, calculation of totals are underestimated. Essentially, zeroes are implicitly imputed for the missing items.

- One method used to deal with unit-level missing data is weighting adjustments. All cases, including the missing cases, are put into classes using variables known for both types. Within the classes, the weights for the missing cases are evenly distributed among the non-missing cases.

- "Imputation" is a process that substitutes values for missing or inconsistent reported data. Such substitutions may be strongly implied by known information or derived as a statistical estimate.

- If imputation is employed and flagged, users can either use the imputed values or deal with the missing data themselves.

- The impact of missing data for a given estimate is a combination of how much is missing (often known via the missing data rates) and how much the missing differ from the sources that provided data in relation to the estimate (usually unknown).

  > For example, given a survey of airline pilots that asks about near-misses they are involved in and whether they reported them, it is known how many of the sampled pilots did not respond. You will not know if the ones who did respond had a lower number of near-misses than the ones who did not.

- For samples with unequal probabilities, weighted missing data rates give a better indication of impact of missing data across the population.

Guidelines

- Unit nonresponse should normally be adjusted by a weighting adjustment as described above.

- Imputing for missing item-level data (see definition above) should be considered to mitigate bias. A missing data expert should make or review decisions about imputation. If imputation is used, a separate field containing a code (i.e., a flag) should be added to the imputed data file indicating which variables have been imputed and by what method.

- The simplest form of imputation is logical imputation where other data collected in the same case, past data, or administrative data imply the correct missing value with near certainty. This method should be used for imputation if available.

  Example: On a travel survey, if the distance of a trip is left blank, but the same person made the same trip more than once, the distance value could be imputed directly from the other trip.

- If a logical method of imputation is not available, then a statistical imputation method of estimation can be applied (assuming, of course, imputation has been deemed appropriate).

  Example: The initial FARS data will have missing the blood alcohol concentration (BAC) data for some people involved in fatal accidents. Past studies showed that leaving this data missing produces seriously biased estimates (for example, crashes with obvious signs of alcohol are usually tested). Therefore, NHTSA developed a statistical model, using other data from the crash to estimate the missing BAC.

- The method of imputation or weight adjustment should be fully documented and summarized in the data system's source and accuracy statement. Imputed fields should be identifiable by some method to help evaluate the impact of imputation, which should also be reported in the source and accuracy statement.

- The missing data effect should be analyzed. For periodic data collections, it should be analyzed after each collection. For continuous collections, it should analyzed at least annually. As a minimum, the analysis should include missing data rates at the unit and item levels and analysis of the characteristics of the reporters and the non-reporters to see how they differ. For some reporting systems, such as with incidents, missing data rates may not be known. For such cases, estimates or just text information on what is known should be provided.

- For sample designs using unequal probabilities (e.g., stratified designs with optimal allocation), weighted missing data rates should be reported along with unweighted missing data rates.

References

- Chapter 4, *Statistical Policy Working Paper 31, Measuring and Reporting Sources of Error in Surveys*, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, July 2001.
- The American Association for Public Opinion Research. 2000. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Ann Arbor, Michigan: AAPOR.

4.3 Production of Estimates and Projections

Principles

- "Derived" data items are additional case-level data that is either directly calculated from other data collected (e.g., # of days from two dates), added from a separate data source (e.g., the weather on a given date), or some combination of the two (e.g., give the departing and arriving airports, calculating distance from an external source). Deriving data is a way to enhance the data set without increasing respondent burden or significantly raising costs.

- An "estimate" is an approximation of some characteristic of the target group, like the average age, constructed from the data.

- A "projection" is a prediction of an outcome from the target group, usually in the future.

> Examples: The average daily traffic volume at a given point of the Garden State Parkway in New Jersey two years from now. Total airline operations ten years from now.

- Estimates from samples should be calculated taking the sample design into account. The most common way this is done is weighted averages using weights based on the design.

- Estimates of standard error of an estimate will give an indication of the precision of the estimate. However, it will not include a measure of bias that may be introduced by problems in collection or design.

Guidelines

- Use derived data to enhance the data set without additional burden on data suppliers.

> For example, the data collection can note the departure and arrival airports, and the distance of the flight can be added derived from a separate table.

- Weights should be used in all estimates from samples. Weights give the number of cases in the target group that each case represents, and are calculated as the inverse of the sampling probability. If using weights, adjust weights for nonresponse as discussed in section 4.2.

> For example, the National Household Travel Survey is designed to be a sample representing the households of the United States, so the total of the weights for all sample households should equal the number of households in the United States. Due to sampling variability, it won't. Since we have a very good count of households in the United States from the 2000 Census, we can do a ratio adjustment of all weights to make them total to that count.

- Construct estimation methods using published techniques or your own documented derivations appropriate for the characteristic being estimated. Forecasting experts should be consulted when determining projections.

> Example: You have partial year data and you want to estimate whole year data. A simple method is to use past partial year to whole year ratios (if stable year to year) to construct an extrapolation projection (Armstrong 2001).

- Standard error estimates should accompany any estimates from samples. Standard errors should be calculated taking the sample design in account. For more complex sample designs, use replicated methods (e.g., jackknife, successive differences) incorporating the sample weights. Consult with a variance estimation expert.

- Ensure that any statistical software used in constructing estimates and their standard errors use methods that take into account the design of the data collection.

- The methods used for estimations and projections should be documented and clearly posted with the resulting data.

References

- Armstrong, J.S. (2001). "Extrapolation of Time Series and Cross-Sectional Data," in *Principles of Forecasting: A Handbook for Researchers and Practitioners*, edited by J. S. Armstrong, Boston: Kluwer.

- Cochran, William G.(1977), *Sampling Techniques* (3rd Ed.). New York: Wiley.
- Wolter, K.M. (1985). *Introduction to Variance Estimatio*n. New York: Springer-Verlag.

4.4 Data Analysis and Interpretation

Principles

- Careful planning of complex analyses involving concerned parties will often ensure a successful result. Data analysis starts with questions that need to be answered. Analyses should be designed to focus on answering the key questions rather than showing all data results from a collection.

- Analysis methods are designed around probability theory allowing the analyst to separate indications of information from uncertainty.

- For analysis of data collected using complex sample designs, such as surveys, the design must be taken into account when determining data analysis methods (e.g., use weights, replication for variances).

- Estimates from 100% data collections do not have sampling error, though they are usually measuring a random phenomenon (e.g., highway fatalities), and therefore have a non-zero variance.

- Data collected at sequential points in time often require analysis with time series methods to account for inter-correlation of the sequential points. Similarly, data collected from contiguous geographical areas require spatial data analysis.

> Note: Methods like linear regression assume independence of the data points, which may make them invalid in time and geographical cases. The biggest impact is in variance estimation and testing.

- Interpretation should take into account the stability of the process being analyzed. If the analysis interprets something about a process, but the process has been altered significantly since the data collection, the analysis results may have limited usefulness in decision making.

- The "robustness" of analytical methods is their sensitivity to assumption violation. Robustness is a critical factor in planning and interpreting an analysis.

Guidelines

- The planning of data analysis should begin with identifying the questions that need to be answered. For all but simplistic analyses, a project plan should be developed. Subject matter experts should review the plan to ensure that the analysis is relevant to the questions that need answering. Data analysis experts should review the plan (even if written by one) to ensure proper methods are used. Even "exploratory analyses" should be planned.

- All statistical methods used should be justifiable by statistical derivation or reference to statistical literature. The analysis process should be accompanied by a diagnostic evaluation of the analysis assumptions. The analysis should also include an examination of the probability that statistical assumptions will be violated to various degrees, and the effect such violations would have on the conclusions. All methods, derivations or references, assumption diagnostics, and the robustness checks should be documented in the plan and the final report.

  Choices of data analysis methods include descriptive statistics for each variable, a wide range of graphical methods, comparison tests, multiple linear regression, logistic regression, analysis of variance, nonparametric methods, nonlinear models, Bayesian methods, control charts, data mining, cluster analysis, and factor analysis (this list is not meant to be exhaustive and should not be taken as such).

- Any analysis of data collected using a complex sample design should incorporate the sample design into the methods via weights and changes to variance estimation (e.g., replication).

- Data analysis for the relationship between two or more variables should include other related variables to assist in the interpretation. For example, an analysis may find a relationship between race and travel habits. That analysis should probably include income, education, and other variables that vary with race. Missing important variables can lead to bias. A subject matter expert should choose the related variables.

- Results of the analysis should be documented and either included with any report that uses the results or posted with it. It should be written to focus on the questions that are answered, identify the methods used (along with the accompanying assumptions) with derivation or reference, and include limitations of the analysis. The analysis report should always contain a statement of the limitations including coverage and response limitations (e.g., not all private transit operators are included in the National Transit Database; any analysis should take this into account). The wording of the results of the analysis should reflect the fact that statistically significant results are only an

indication that the null hypothesis may not hold true. It is not absolute proof. Similarly, when a test does not show significance, it does not mean that the null hypothesis is true, it only means that there was insufficient evidence to reject it.

- Results from analysis of 100 percent data typically should not include tests or confidence intervals that are based on a sampling concept. Any test or confidence interval should use a measure of the variability of the underlying random phenomenon.

> For example, the standard error of the time series can be used to measure the variance of the underlying random phenomenon with 100 percent data over time. It can also be used to measure sampling error and underlying variance when the sample is not 100 percent.

- The interpretation of the analysis results should comment on the stability of the process analyzed.

> For example, if an analysis were performed on two years of airport security data prior to the creation of the Transportation Security Agency and the new screening workforce, the interpretation of the results relative to the new processes would be questionable.

References

- Skinner, C., D. Holt, and T. Smith. 1989. *Analysis of Complex Surveys*. New York, NY: Wiley.
- Tukey, J. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- Agresti, A. 1990. *Categorical Data Analysis*. New York, NY: Wiley.

5. Dissemination of Information

"Dissemination means agency initiated or sponsored distribution of information to the public. Dissemination does not include distribution limited to government employees or agency contractors or grantees; intra- or inter-agency use or sharing of government information; and responses to requests for agency records under the Freedom of Information Act, the Privacy Act, the Federal Advisory Committee Act, or other similar law. This definition also does not include distribution limited to correspondence with individuals or persons, press releases, archival records, public filings, subpoenas or adjudicative processes." – OMB Guidelines.

The first key point in disseminating statistical information is the principle of openness relative to all aspects of quality. Pursuant to that principle, the statistical information being disseminated must be accompanied by documentation. That documentation can be separately posted documents referenced by the disseminated information or it can be part of the disseminated entity. The second key point in the dissemination is the final reviews before dissemination. These quality reviews are a final assurance that all quality control steps have been taken and that the dissemination package is complete.

5.1 Publications and Disseminated Summaries of Data

Principles

- In publications or summaries, information should be clearly presented to users, and users should be informed about the source(s) of the information presented.

- As far as possible, tables, graphs, and figures should be interpretable as stand-alone products in case they become separated from their original context.

- Methods used to produce data displayed in tables, graphs, and summary data should be available to the reader.

- Statistical interpretations should indicate the amount of uncertainty.

Guidelines

- Documents should be well organized with language that clearly conveys the message intended. Tables, graphs, and figures should be consistent with each other and the text discussing them.

- All tables, graphs, figures that illustrate data, and text that provides data not in accompanying illustrations should include a title. Titles for tables and graphs should be clearly worded and answer three questions: what (data presented), where (geographic area represented), and when (date covered by data).

- All tables, graphs, figures that illustrate data, and text that provides data not in accompanying illustrations should include a source reference. The source reference should contain one or more entries with references to the sources for the information presented. The reference should be sufficiently detailed for a reader to locate the data used. Since databases and documents may be updated, the "as of" date for the source should also be noted.

- Footnotes should be used, if necessary, to clarify data illustrations, tables, graphs, and figures to clarify particular points, abbreviation symbols, and general notes.

- The style of a publication should conform to specific agency style guidelines to ensure consistency and clarity throughout the document.

- Documents disseminated on the Internet should be accessible as required by section 508 of the Rehabilitation Act (29 USC 794d).

- A contact point should be provided in the publication or with the summaries to facilitate user comments and suggestions.

- Documents containing estimates, projections, and analyses should contain or reference the methodology supporting documentation required in sections 4.3 and 4.4.

References

- U.S. Government Printing Office Style Manual

5.2 Micro data Releases

Principles

- The term "micro data" refers to data files with various information at the "unit" level. The unit is dependent upon what data are being collected and from what sources.

> Examples: micro data may be a collection of individual responses from each person or each household to a survey, reports of information from each company, or reports of individual incidents.

- Making micro data available can enhance the usefulness of the information, and can assist the public in determining whether results are reproducible. However, micro data should not be released in violation of existing protections of privacy, proprietary information, or confidentiality.

- Micro data should be provided in a manner that facilitates its usefulness to users.

- Quality information as recommended herein, file layouts, and information describing the data (i.e., metadata) enhance the usefulness of the micro data.

Guidelines

- Micro data released to the public should accessible by users with generally available software. It should not be restricted to a single application format.

- Micro data should be accompanied (or have a reference to) by the quality-related documentation discussed herein: planning documentation and collection, processing, and analysis methodology.

- Microdata releases should be accompanied by file layouts and information describing the data.

- Micro data should be accompanied by a clear description of revision information related to the file.

- A contact point should be provided with the data to facilitate user comments and suggestions.

References

- International Standardization Organization standard 11179, Specification and Standardization of Data Elements

5.3 Source and Accuracy Statements

Principles

- Source and Accuracy Statements (S&As) are compilations of data quality information discussed herein. They provide information on where the data came from, how it was collected, and how it was processed. They include information on known strengths and weaknesses of the data.

- S&As should be regularly updated to include changes in methodology and results of any quality assessment studies.

Guidelines

- The S&A for a data source should contain or refer to the current data system objectives and data requirements as discussed in sections 2.1 and 2.2 of these guidelines.

- The S&A for a data source should contain the data source and data collection design as discussed in section 2.3 and 2.4 of these guidelines.

- The S&A for a data source should contain or refer to the collection operations methodology documentation discussed in section 3.2 and 3.3.

- The S&A for a data source should contain or refer to the processing documentation discussed in sections 4.1 – 4.4.

- The S&A for a data source should describe major sources of error including coverage of the target population, missing data information, measurement error, and error measures from processing quality assurance.

- The S&A for a data source should contain or reference the revision process for the system and should indicate the source for revision information for the data source.

References

- General Accounting Office, *Performance Plans: Selected Approaches for Verification and Validation of Agency Performance Information*, GAO/GGD-99-139 (July 1999).
- Office of Management and Budget, *Statistical Policy Working Paper 31: Measuring and Reporting Sources of Error in Surveys* (July 2001).

5.4 Pre-Dissemination Reviews

Principles

- Informal and formal reviews of publications, summaries, or micro data will help ensure that a data product meets a minimal level of quality.

- Due to the diverse aspects of quality in a final product, reviews need to be conducts by several people with different backgrounds.

- Reviews of documentation produced through the various stages of data development will enhance the review process.

Guidelines

- A subject matter specialist other than those directly involved in the data collection and analysis should review the plans, methodology documents, and reports prior to dissemination. They should also review publications and summaries resulting from the data for content and consistency.

- Publications should be reviewed by style and visual information specialist for compliance with style standards.

- A statistician or other data analysis specialist other than those directly involved in the data collection and analysis should review the plans, methodology documents, and reports prior to dissemination for compliance with these guidelines. They should also review publications and summaries resulting from the data for the wording of statistical interpretation.

- Any items to be disseminated via the Internet should be reviewed by a Section 508 compliance specialist for accessibility.

- Any data products that will be disseminated via special software onto the Internet should be tested for accessibility and interpretability prior to dissemination.

- For micro data releases, the release files and the metadata should be reviewed by an information technology specialist for clarity and completeness.

- If an external peer review process is used: (1) peer reviewers should be selected primarily on the basis of necessary technical expertise; (2) peer reviewers should be expected to disclose to DOT prior technical/policy positions they may have taken on the issues at hand and their sources of personal and institutional funding (private or public); and (3) peer reviews be conducted in an open and rigorous manner.

References

- Ott, E., E. Shilling, and D. Neubauer. 2000. Process Quality Control: Troubleshooting and Interpretation of Data. New York, NY: McGraw-Hill.

6. Evaluating Information Quality

Once a data system exists, the key to achieving and maintaining a high level of data quality is to regularly assess all aspects of data quality and improve the data collection and processing system. That can be accomplished by regular assessments of the data collected, special studies of aspects of the data and the effectiveness of the collection and processing processes, and quality control of key processes to both control the quality during operation and to collect data quality information.

6.1 Data Quality Assessments

Principles

- "Data quality assessments" are data quality audits of data systems and the data collection process.
- Data quality assessments are comprehensive reviews of the data system to note to what degree the system follows these guidelines and to assess sources of error and other potential quality problems in the data.
- The assessments are intended to help the data system sponsor to improve data quality.
- The assessments will conclude with recommendations for data quality improvements.

Guidelines

- Since data users do not have the same access to or exposure to information about the data system that its sponsors have, the data system sponsors should make the initial data quality assessment.
- Data quality assessments should be undertaken periodically to ensure that the quality of the information disseminated meets requirements.
- Data quality assessments should be used as part of a data system redesign effort.
- Data users, including secondary data users, should be consulted to suggest areas to be assessed, and to provide feedback on the usefulness of the data products.
- Assessments should involve at least one member with a knowledge of data quality who is not involved in preparing the data system information for public dissemination.

- Findings and results of a data quality assessment should always be documented.

References

- General Accounting Office, *Performance Plans: Selected Approaches for Verification and Validation of Agency Performance Information*, GAO/GGD-99-139 (July 1999).

6.2 Evaluation Studies

Principles

- Evaluation studies are focused experiments carried out to evaluate some aspect of data quality.

- Many aspects of data quality cannot be assessed by examining end-product data.

- Evaluation studies include re-measurement, independent data collection, user surveys, collection method parallel trials (e.g., incentive tests), census matching, administrative record matching, comparisons to other collections, methodology testing in a cognitive lab, and mode studies.

- "Critical data systems" are systems that either contain data identified as "influential" or provide input to DOT-level performance measures.

Guidelines

- Critical data systems should have a program of evaluation studies to estimate the extent of each aspect of non-sampling error periodically and after a major system redesign.

- Critical data systems should periodically evaluate bias due to missing data, coverage bias, measurement error, and user satisfaction.

- All data systems should conduct an evaluation study when there is evidence that one or more error sources could be compromising key data elements enough to make them fail to meet data requirements.

- All data systems should conduct an evaluation study if analysis of the data reveals a significant problem, but the source is not obvious.

References

- General Accounting Office, *Performance Plans: Selected Approaches for Verification and Validation of Agency Performance Information*, GAO/GGD-99-139 (July 1999).
- Office of Management and Budget, *Statistical Policy Working Paper 31: Measuring and Reporting Sources of Error in Surveys* (July 2001).
- Lessler, J. and W. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York, NY: Wiley.

6.3 Quality Control Systems

Principles

- Activities in survey collection and processing will add error to the data to some degree. Therefore, each activity need some form of quality control system to prevent and/or correct error introduced during the activity.

- The more complex or tedious an activity is, the more likely error will be introduced, and therefore, the more elaborate the quality control needs to be.

- A second factor that will determine the level of quality control is the importance of the data being processed.

- Data system activities that need extensive quality control are check-in of paper forms, data entry from paper forms, coding, editing, and imputation.

- Quality control methods include 100% replication, as with key entry of critical data, sample replication (usually used in a stable continuous process), analysis of the data file before and after the activity, and simple reviews.

Guidelines

- Each activity should be examined for its potential to introduce error.

- The extent of quality control for each activity should be based on the potential of the activity to introduce error combined with the importance of the data.

- Data should be collected from the quality control efforts to indicate the effectiveness of the quality control and to help determine whether it should be changed.

- The quality control should be included in the documentation of methods at each stage.

References

- Ott, E., E. Shilling, and D. Neubauer. 2000. Process Quality Control: Troubleshooting and Interpretation of Data. New York, NY: McGraw-Hill.

6.4 Data Error Correction

Principles

- No data system is free of errors.

- Actions taken when evidence of data error comes to light are dependent on the strength of the evidence, the impact that the potential error would have on primary estimates produced by the data system, and the resources required to verify and correct the problem.

Guidelines

- A standard process for dealing with possible errors in the data system should exist and be documented.

- If a disseminated data file is "frozen" for practical reasons (e.g., reproducibility and configuration management) when errors in the data become known, the errors should be documented and accompany the data.

References

- Ott, E., E. Shilling, and D. Neubauer. 2000. Process Quality Control: Troubleshooting and Interpretation of Data. New York, NY: McGraw-Hill.